MEDIX: The Human Health Analyzer

Priyanshi Sharma[1], Sheetal Agarwal[2], Tushank Garg[3], Amit Saini[4], Dr. Mukesh Rawat[5]
[1,2,3]*Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut 250005, U.P., India*
[4]*Guide, Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut 250005, U.P., India*
[5]*Supervisor, Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut 250005, U.P., India*

Abstract **- Data mining and machine learning algorithms play a key role in this area. Researchers use many data processing methods to assist the field of health care in predicting diseases. However, a proposed classification approach based on Random Forest Classification (RF) to assign each data sample to its appropriate class. ML algorithms can also be used to find correlations and associations between different diseases. Our system uses KNN for diabetes and RF for heart & breast cancer detection.**

Index Terms **- Classification, Heart disease, diabetes and breast cancer machine learning, KNN algorithm, random forest algorithm.**

INTRODUCTION

Prediction using machine learning methods is the key focus. Machine learning is now widely used in many business applications such as e-commerce and many more for a few days. Prediction is one of the fields in which this machine learning is used, the subject of prediction of heart disease, diabetes & breast cancer processing user data set and user data to which we need to predict the probability of heart disease or diabetes or breast cancer incidence.

Healthcare field today generates large amounts of complex data about patients. However, to diagnosing heart disease, diabetes& breast cancer, several time-consuming tests and analyzing critical factors are done. Doctors can often fail to make specific decisions when diagnosing a patient's heart disease, breast cancer or breast cancer, so health disease prediction systems that use machine learning algorithms help to achieve precise results in those cases. We especially focused on important attributes like, high blood pressure, abnormal blood lipids, cholesterol, chest pain level, age, gender, family generation, etc to predict whether person is suffering with heart disease or diabetes or breast cancer not. Diabetes mellitus is a condition characterized by a metabolic disorder and an irregular increase in blood sugar levels caused by or attributed to insulin deficiency or low tissue sensitivity to insulin. Heart disease is really a headache for all doctors all over the world.

Healthcare industry now a days, generates large amount of data of patients.

Research objective: The objective of this research is to undergo a comparative study on various decision tree classifier algorithms and to identify the best classifier for heart disease or diabetes or Breast cancer classification.

METHODOLOGY

A. Machine Learning:

Machine learning is ability of computer to learn make decisions or classify objects by itself. As like humans machine also gets ability to think and take appropriate decisions.

ML model is said to learn from experience E. Experience comes from the dataset it gets training from. So, to create a better model, we need to provide a better & large set of data.

There are 2 types of machine learning techniques, and we are using supervised machine learning techniques for detecting disease.

Classification algorithms

These are machine learning algorithms that classifies. objects or labels from multiple inputs. There are many classification algorithms in the field of machine

learning, we are using Random Forest and K- Nearest Neighbor.

B. Random Forest (RF):

Random Forest is a supervised machine learning algorithm that is used for classifying purpose. The word forest is used because, we use multiple decision tree's output for final decision.

Advantages of RF:

- It is versatile in nature as it can be used for both, classification and regression.
- Rf are more robust than a single decision tree as many trees are playing role in this algorithm for taking decision.

Algorithm Random Forest:

Steps of Random Forest classifier are shown in figure 1.

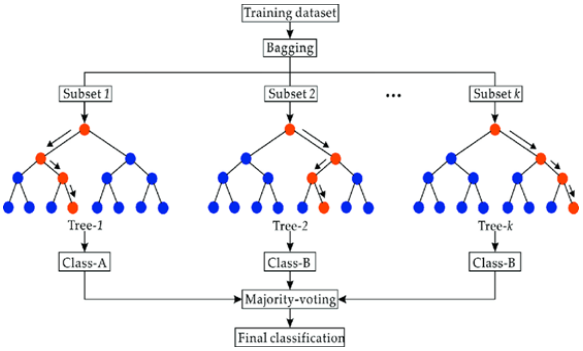

Fig 1. RF-algorithm

C. K-NEAREST NEIGHBOR(KNN):

One of the simplest machine learning algorithms for classification is the KNN algorithm. One of the simplest machine learning algorithms for classification is the KNN algorithm. The aim of KNN is to use a dataset in which data points are segregated for the prediction of classification into several groups. The figure below provides an example of the KNN algorithm.

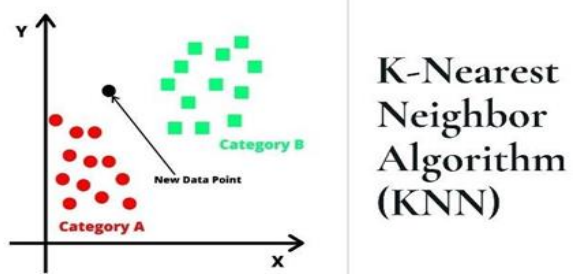

Fig 2.KNN example

RESULT

Experiments were carried out using the Cross-validation method. The accuracy of the disease dataset comparison is shown in Table 1 and Figure 1. After ten-cross validation, the results are obtained. Our solution is 7.97 percent better than the algorithm C4.5.

We contrasted DT with our approach, the outcome is shown in Table 2 and Figure2. Our approach achieved 100 percent precision, and 98.66 percent was obtained by DT. Various parameters are compared and described in Table 3 and Figure 3 for the T.S disease data collection.

The experimental results above indicate that our RF method was better than other algorithms, so we trained our model using the classification algorithm RF (Random Forest).

Table 1: Comparison of Precision for Diseases Data Collection.

Sl.no	***Approach***	***Accuracy***
1	PART C4.5	75.73
2	Na ïve bayes	78.56
3	Decision table	82.43
4	Neural nets	82.77
5	**Our approach**	**83.70**

Table 2: Comparison of Precision for Diseases Data Collection.

Sl.no	*Approach*	*Accuracy*
1	Decision Tres(DT)	98.66
2	Our approach	100

Table 3: Comparison of different parameters for the T.S. Diseases Data Set.

Sl.no	*Parameter*	*Our approach*	*Decision tree*
1	Sensitivity	100	100
2	Specificity	100	92.86
3	Disease prevalence	82.67	81.33
4	Positive Predictive Value(PPV)	100	98.39
5	Negative Predictive Value(NPV)	100	100

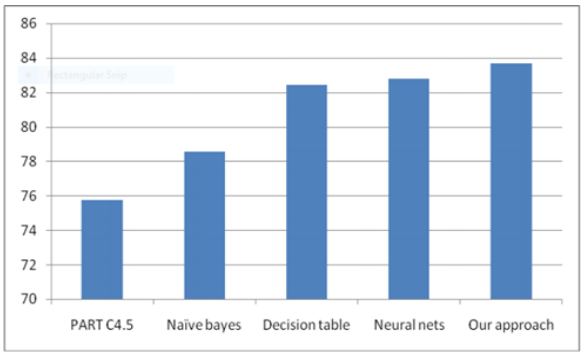

Figure 1: Comparison of accuracy of heart stalog data set across different approaches.

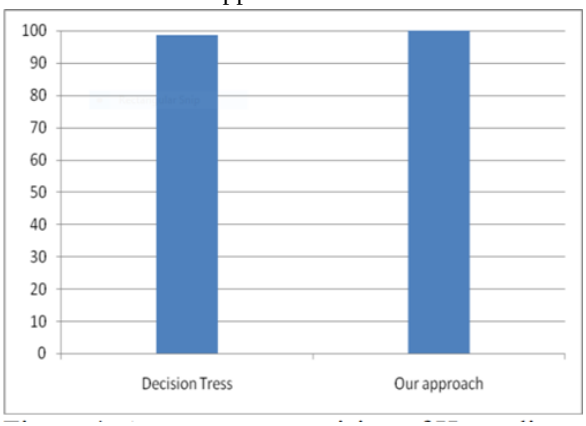

Figure2: Comparison of Heart Disease Data Set-T. SS Accuracy Comparison

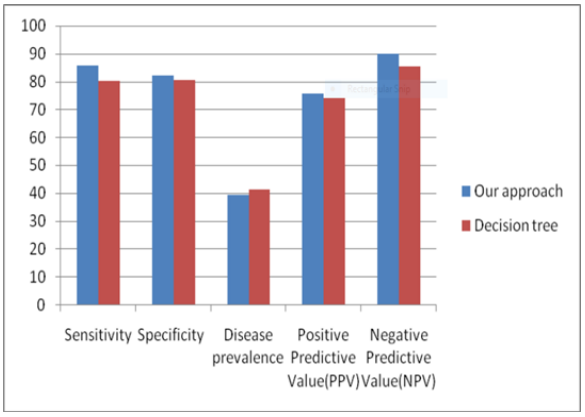

Figure 3: Comparison of different Heart Disease Data Set-T.S. parameters.

CONCLUSION

In this research paper, we have established an effective and precise way to use Random Forest for disease detection. Algorithm has shown 0.763 precision and 0.930 accuracy, this proves that RF is best among all other algorithms to detect the diseases. The algorithm has shown 75-80% accuracy in predicting the class label of unknown records. Our approach proved to be better than traditional classification algorithms for effective classification of heart disease. These machine learning algorithms can be used to predict many disease prediction or detection like heart attack, asthma, diabetes and high blood pressure etc.

REFERENCE

[1] https://youtu.be/D_2LkhMJcfY
[2] https://www.sciencedaily.com/releases/2016/06/1606231 15738.htm
[3] https://link.springer.com/article/10.1007/s42979-020- 00365-y
[4] https://medium.com/@enfageorge/predicting-breast-cancer-using-random-forest-classifier-d193c72de8a3
[5] https://www.researchgate.net/publication/331264492_Using_Random_Forest_Algorithm_for_Breast_Cancer_ Diagnosis
[6] https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[7] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
[8] https://www.acls.net/a-guide-to-a-healthy-heart.htm
[9] https://www.webmd.com/diabetes/guide/default.htm
[10] https://www.nationalbreastcancer.org/breast-health- guide/